Step 0 (Initialization):
\( D \) = the set of the documents to be clustered
\( D_m \) = the set of documents in \( D \) indexed with a given MeSH term \( m \) (for each \( m \) in \( M \))
\( N_m \) = the number of documents in \( D \) indexed with a given MeSH term \( m \) (for each \( m \) in \( M \))
\( M \) = the set of MeSH terms collected from all documents in \( D \) (after removing the terms that are given on a stoplist of the top 20 most frequent in MEDLINE, and removing each term \( m \) for which the relative document frequency \( |D_m|/|D| \) is > 1/3)
\( i = 0 \) (cluster number)

Step 1 (Iterations):
\[ \text{WHILE } (i = i + 1 < 15 \text{ and } M \neq \emptyset) \{ \]
\( L_i = m \) in \( M \): \( N_m \geq N_n \) for all \( n \) in \( M \) (identifies the label for the \( i \)-th cluster)
\( C_i = D_m \) (assigns a set of papers to the \( i \)-th cluster \( C_i \))
\( D = D - D_m \) (removes the papers in the \( i \)-th cluster from being considered for the potential “Miscellaneous” cluster, and from contributing to the counts of the remaining MeSH terms)
\( N_n = \text{number of remaining documents in } D \text{ with a given MeSH term } n \) (for each \( n \) in \( M \))
\( M = M - m \) and its children* (removes the \( i \)-th cluster label and its children from further consideration)
\[ \}
\] IF (\( D \neq \emptyset \)) {
\( i = i + 1 \)
\( L_i = \text{“Miscellaneous”} \)
\( C_i = D \)
}\n\( n = i \)

Step 2 (Output Clusters): List of cluster labels \( L_i \)'s each with a corresponding set of documents \( C_i \)'s for \( i = 1, 2, \ldots, n \), displayed in order of decreasing size of the document clusters.

*Given two MeSH terms \( m \) and \( n \), \( m \) is considered a child of \( n \) if \( m \) occurs below \( n \) in the MeSH hierarchy.

Notice that the counts (\( N_m \)'s) are reduced during the execution of the WHILE loop, whereas the document sets (\( D_m \)'s) stay the same. Thus, each document may be assigned to multiple clusters because cluster labels are chosen based on the counts (\( N_m \)'s) and the clusters are based on document sets (\( D_m \)'s).